MBMG
Metabarcoding & Metagenomics

Metabarcoding and Metagenomics 8: 277–294 (2024)
DOI: 10.3897/mbmg.8.124564

Software Description

The ASV Registry: a place for ASVs to be

Christian Bräunig[1], Sandra Meid[1], Björn Quast[1], Vera Rduch[2], Peter Grobe[1]

1 *Department of Biodiversity Informatics, Leibniz Institute for the Analysis of Biodiversity Change, Adenauerallee 127, Bonn, Germany*
2 *Center for Taxonomy and Morphology, Leibniz Institute for the Analysis of Biodiversity Change, Adenauerallee 127, Bonn, Germany*
Corresponding author: Christian Bräunig (c.braeunig@leibniz-lib.de)

Abstract

Despite the effectiveness of DNA metabarcoding for gaining insights into biodiversity and environmental species composition, a centralised management and storage option including easy accessibility of already-published data is lacking. Since most data are published as supplementary material or in private repositories, DNA metabarcoding has a huge untapped potential to be used for analyses across multiple taxa, sample locations or multiple research projects. We developed a platform to register, manage and identify amplicon sequence variants (ASVs) or zero-radius OTUs (ZOTUs), respectively, against several barcode reference datasets. Moreover, ASV tables can be uploaded, managed, versioned and published with DOIs thus contributing to the full research Data Life Cycle.

Key words: Amplicon sequence variant, barcode reference databases, biodiversity monitoring, data storage, DNA metabarcoding, operational taxonomic unit, taxon annotation

OPEN ACCESS

This article is part of:
Towards standardized molecular biodiversity monitoring
Edited by Teresita Porter, John Darling, Kelly Goodwin, Tiina Laamanen, Kristian Meissner, Toshifumi Minamoto

Academic editor: Toshifumi Minamoto
Received: 11 April 2024
Accepted: 23 August 2024
Published: 4 October 2024

Citation: Bräunig C, Meid S, Quast B, Rduch V, Grobe P (2024) The ASV Registry: a place for ASVs to be. Metabarcoding and Metagenomics 8: e124564. https://doi.org/10.3897/mbmg.8.124564

Copyright: © Christian Bräunig et al.
This is an open access article distributed under terms of the Creative Commons Attribution License (Attribution 4.0 International – CC BY 4.0).

Introduction

DNA metabarcoding allows for the routine identification of species and analysis of species composition in environmental samples. If well-curated reference data are available, this method can yield comparable results to or even outperform conventional species identification by human experts in quantity, quality and speed (Elbrecht et al. 2017; Remmel et al. 2024). This is also why DNA metabarcoding is not only an important tool for science, but is also increasingly used by environmental authorities (Leese et al. 2023). Hence, it is reasonable to consider it a future standard method in any biological research where species occurrence and distribution are of interest, such as ecological research and biomonitoring (Porter and Hajibabaei 2018).

The sequencing reads produced by DNA metabarcoding usually have two possible fates: individual reads may be clustered based on sequence similarity, thereby generating molecular operational taxonomic units (OTUs; Blaxter et al. (2005); Callahan et al. (2017)). Alternatively, single sequences are extracted and quality assured from multiple reads by bioinformatics tools and become amplicon sequence variants (ASVs; Callahan et al. (2017)). As single DNA sequences, they offer a finer resolution and a higher degree of comparability across datasets (Callahan et al. 2017). It is worth pointing out that the terms ESV (Exact Sequence Variant) and ZOTU (Zero-radius OTU) are used

interchangeably with ASV and refer to the same concept (Antich et al. 2021). In favour of simplicity, solely ASV will be used hereafter.

In case multiple environmental samples were analysed simultaneously, organising the detected sequences in an ASV table is reasonable (Callahan et al. 2017). Such a table combines the extracted sequences of all sampling plots with the number of occurrences of each sequence within the respective plots. It can be enriched by taxonomic identification of the individual sequences. The production of these taxonomic assignments is the step at which the species composition of the samples is revealed. This can be done using tools like BLAST (Camacho et al. 2009), the Barcode Of Life Data (BOLD; Ratnasingham and Hebert (2007)) system identification engine or VSEARCH (Rognes et al. 2016) against DNA barcode reference databases like the BOLD database or the German Barcode of Life library (GBOL; German Barcode of Life Consortium (2011)). Some of the analysed sequences will have no species or genus name associated because either they stem from DNA of specimens that belong to species not contained in the reference database or they belong to distinct populations that are genetically different from the reference material of the same species in the database. The resolution of species detection hinges on the quality and coverage of the used reference libraries (Macher et al. 2021). Since the number of available sequences and the quality of identifications in reference databases are growing over time (Weigand et al. 2019), the number and precision of taxon assignments will increase continuously. The introduction of new marker sequences and improvements in search tools will further lead to improved taxon assignments. Thus, the taxon assignments in ASV tables are subject to constant change. Repeated taxon identification of a given set of ASV sequences is, hence, not only worthwhile to make use of this ongoing growth and improvement, but necessary to keep the identifications up to date and as informative as possible.

While DNA metabarcoding can quickly generate large amounts of data, most ASV tables are still stored as supplements to publications or in private repositories. This greatly impairs access and reuse of these data. As a result, analyses across multiple research projects are difficult, error-prone and often not up to date. Efforts, like the EBI metagenomics with MGnify (Richardson et al. 2023) serve the needs for uploading and annotating environmental DNA samples (Mitchell et al. 2017), but there is no registry for ASV tables in which they can be sustainably stored, managed, versioned and published in a well-documented manner.

The ASV Registry presented here contributes to the completion of the Research Data Life Cycle (Rüegg et al. 2014; Griffin et al. 2017) and makes ASV data available to open science under the principles of findability, accessibility, interoperability and reusability (FAIR; Wilkinson et al. 2016).

Implementation

The ASV Registry

To address the need for centralised archival and management of ASV data, we have developed the ASV Registry. This platform allows users to archive, analyse, manage and retrieve ASV data in consideration of the FAIR data principles (Wilkinson et al. 2016). We have designed the ASV Registry as a full-fledged web application to impose minimal software requirements on users. It is reachable

via any browser at https://asv.bolgermany.de/metabarcoding and full usage only requires creating an account. The services provided are entirely free of charge.

Technically, the web platform is based on the Pyramid web framework (https://trypyramid.com/) and sits atop a MySQL database (https://www.mysql.com/). An ElasticSearch indexer (https://www.elastic.co/elasticsearch) manages search and retrieval processes.

Data life cycle and key features

The ASV Registry supports a specific life cycle for ASV tables to facilitate effective and efficient usage of the contained data. Before the upload of any data, users must first create or be added to a Project container (Fig. 1). Analogous to real projects, these containers act as organisational units to group uploaded data and support collaborative work, while also handling access management. The amount of data that these containers may hold is not limited. Any number of ASV tables may be uploaded to any one project. We envision the project containers to hold all ASV datasets produced within a given real-life research project. Only members of a given project can upload, analyse, delete and publish the data within it. Project creators can dynamically add and remove members.

For upload, ASV tables must be in CSV, TSV or XLSX format and must feature a basic layout (Table 1). This layout consists of a single column of DNA sequences, corresponding to each ASV, as well as one or more columns, each representing a single sampling plot or event. Cells within the sample columns contain the occurrence data as integers. User-defined identifiers for each ASV sequence can also be included in the table. Further data may be present in the uploaded file; however, these will not be written into the database. Sequences, occurrence data and identifiers are exclusively read and saved in the database during upload (See Appendix 1: Fig. A1 for a detailed description of the database schema and backend handling). The uploaded file is saved in its entirety on the server. During download, it serves as a template that is filled with the identifiers and annotation data generated in the ASV Registry. Collection metadata enclosed in an ASV table file should be copied to a separate file and uploaded as a Collection table (Fig. 2).

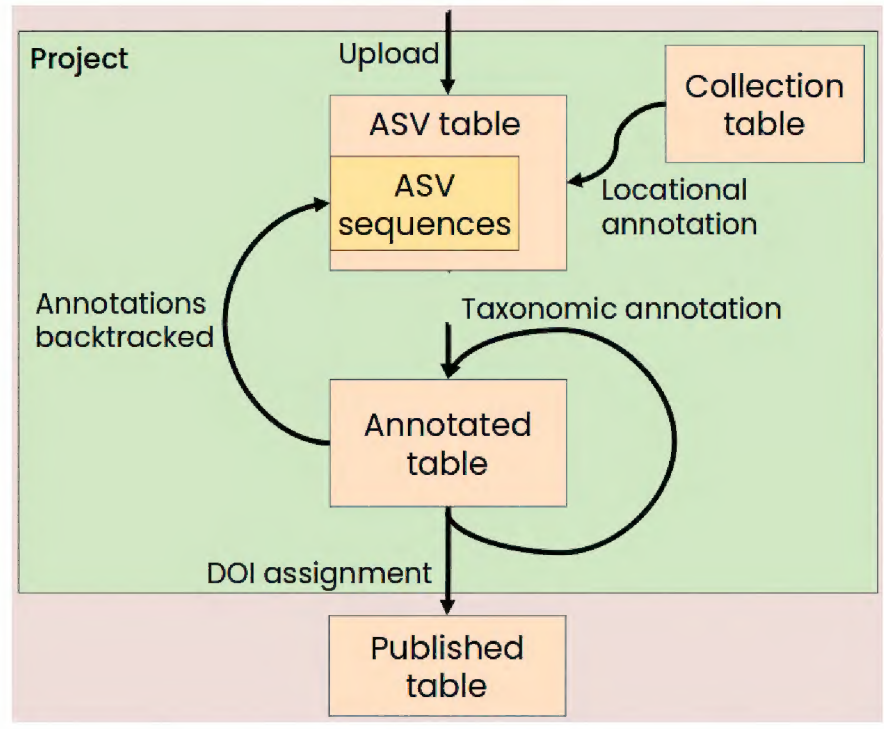

Figure 1. ASV table-specific data life cycle. The Project container (green) acts as a collaborative workspace and access management. Publication of a table via DOI assignment moves the table out of this container and makes it publicly available.

Table 1. Initial data format. ASV tables that are to be uploaded must at the very least contain the following information: the DNA sequence of each ASV (green) and the integer-based read count of each ASV per sampling plot (blue), where each column corresponds to one sample. User-defined identifiers for the ASVs (orange) can also be included. These will be saved in the database, but will not be used as identifiers during later processing.

asv_id	PCRPCR1C	PCRPCR1a	PCRPCR1b	PCRPCR2C	PCRPCR2a	PCRPCR2b	Seq
uniq152440	1	3	1	0	1	0	TCTATCTTCATACTTATTTCATTCTTCTCCAT-CAATTGATATTGCAATTTTTTCAATACATATTACAG-GAATTTCCTCTGTTATTGGTTCTTTAAATTTTATTGTA-ACTATTATATTTATAAAAAATTTTTCCTTAAATTATGAT-CAAATCAATTTATTTTCTTGATCTGTATGTATTACAG-TAATTCTACTAATTTTATCTTTACCAGTTTTAGCAG-GAGCAGTAACTATACTTCTATTTGATCGAAATTTTA-ATACATCCTTTTTTGATCCTATAGGAGGAGGTGATC-CAATTCTTTATCAACATTTATTC
uniq152443	1	0	3	0	1	1	TTTATCATCTAGAATTGCTCATGGAGGAGCTTC-CGTAGATTTAGCTATTTTCTCTCTTCACCTAGCAG-GAATTTCTTCAATTCTTGGGGCAGTAAATTTTATTA-CAACTGTAATTAATATACGTTCTACAGGAATTACCTTT-GATCGAATACCTTTATTTGTATGATCTGTAGTTATTA-CAGCTATTTTATTACTTCTTTCTTTACCCGTATTAG-CAGGAGCAATTACTATATTATTAACAGATCGAAATCTA-AATACCTCATTCTTTGACCCAGCAGGAGGAGGAGAC-CCTATCTTATATCAACATTTATTC
uniq152454	3	1	0	0	2	0	ACTTTCATCTAATATTGCCCATGGAGGTAGTTCT-GTAGATCTTGCTATTTTTTCTCTTCATTTAGCTG-GAATTTCTTCTATTCTAGGAGCTATTAATTTTATTAC-TACCATTATTAATATACGATTAAATAGTTTATCTTTT-GATCAAATACCTTTATTTATCTGAGCTGTTGGAATTA-CAGCATTTTTATTATTATTATCTTTACCTGTTTTAG-CAGGAGCTATTACTATATTATTAACTGATCGTAATTTA-AATACATCATTTTTTGACCCTGCAGGAGGGGGAGAC-CCTATTTTATATCAACACCTATTT
uniq152461	2	0	3	0	0	1	ACTATCAAATAATATATCCCATGCAGGAGCATCAG-TAGATTTAACAATTTTTTCCCTTCACTTAGCCG-GAATTTCATCTATTCTAGGAGCTATTAATTTCATTTC-TACAGTAATTAATATACGAACTACAGGAATAAAATTT-GAACGAATACCCTTATTTATTTGAGCTGTTACTTTAA-CAGCAATTCTATTATTATTATCATTACCAGTTTTAGCT-GGAGCGATTACAATATTATTAACTGATCGAAACTTA-AATACTTCATTTTTTGATCCTTCAGGTGGAGGAGAC-CCTATTTTATATCAACATTTATTT
uniq152464	1	3	2	0	0	0	ACTTGCAGGAGCTATTGCTCATGGTGGAGGATCAGTT-GATCTAGCCATTTTTTCATTACATCTAGCCGGTGTCT-CATCAATTTTAGGAGCAGTAAACTTCATTACAACAG-CAATCAACATGCGGTCGGAAAGAATAACACTGGAT-CAAACACCACTATTTGTTTGATCAGTAGCAATCACAG-CACTTCTATTACTACTATCATTACCAGTACTAGCAG-GAGCAATTACTATATTATTAACTGATCGAAATTTA-AATACATCATTCTTCGACCCAGCTGGAGGTGGAGATC-CAATTCTCTATCAACATTTATTT
uniq152471	1	1	3	1	0	0	ATTGTCGAACAATATATTTCACAGAGGTGCCTCAG-TAGACATAGCCATCTTTTCACTACACTTAGCAGGTG-TATCATCTATCTTAGGAGCCATTAACTTTATTTCCA-CAATCATGAACATACGACCCACAGGTATAACACCT-GAACAAATTCCCCTATTCGTATGATCTGTAGGAAT-CACTGCTCTATTATTATTATTATCACTCCCAGTTTTAG-CCGGAGCAATCACTATATTATTAACAGAC-CGAAACTTCAACACATCCTTCTTTGACCCTACGGGGG-GAGGAGATCCTATTCTATACCAACACTTATTT
uniq152480	0	3	1	0	1	1	ACTATCTGCTAATATTGCTCATAGAGGAGCATCAG-TAGATATAGCAATTTTCTCACTACATTTAGCAGGAG-TATCATCAATTCTAGGAGCAGTAAACTTCATTTCAAC-TATTATTAATATACGGCCCGTAGGTATATCACCT-GAACGAATTCCGCTATTTGTATGATCAGTAGGAAT-TACTGCATTACTATTATTATTATCACTACCTGTATT-AGCAGGGGCTATTACAATACTATTAACAGATCGTA-ATTTTAATACATCATTTTTTGACCCAGCCGGAGGTGG-GGATCCTATTTTATATCAACATTTATTC

asv_id	PCRPCR1C	PCRPCR1a	PCRPCR1b	PCRPCR2C	PCRPCR2a	PCRPCR2b	Seq
uniq152481	0	3	1	1	1	0	TCTATCCTCAAATATTGCACATAACGGTTCATCT-GTAGACCTGGCAATTTTCAGATTACATTTAG-CGGGGATCTCCTCAATTCTAGGAGCAGTTA-ATTTTATTTCTTCTATAATTAATATACGACCTTTA-AAAATAAGATTTGATCAAATACCTCTTTTTACTT-GAGCTGTAGCTATTACAGCTCTCT-TACTCCTCCTCTCTCTGCCCGTACTAGCTGGAGCGAT-TACTATACTTCTCACAGATCGAAACATTAATACAT-CATTTTTTGATCCCGCAGGAGGTGGAGATCCTATTC-TATACCAACACTTATTT
uniq152482	0	5	0	0	1	0	ACTTGCAGGAGCTATTGCTCATGGTGGGTGATCAG-TAGATCTAGCCATTTTTACATTACATCTAGCCGGT-GTTTCATCAATTTTAGGAGCAGTAAACTTTATTAAAA-CAGCAATTAATATGCGGTCGGGGAGAATAACACTG-GATCAAACCACTATTTGTCTGATCAGTAGCAATTA-CAGCACTTTTATTACTATTACCATTACCAGTTTTAG-CAGGAGCAATTACTATATTATTAACTGATCGAAACTTA-AATACATCATTCTTTGATCCAGCTGGAGGTGGAGATC-CAATTCTTTATCAACATTTATTC
uniq152483	0	2	2	1	1	0	CCTATCTTCAGGAATTGCACATGGAGGTGCTTCT-GTAGACTTAGCTATTTTTTCTCTACATTTAGCTG-GTGTCTCATCAATTTTAGGAGCAGTAAATTTTAT-TACTACAGTTATTAATATACGATCTACTGGAAT-TACTTTCGATCGAATGCCATTGTTTGTATGAGCAGT-GGTAATTACTGCTTTATTACTTCTCCTTTCTTTACCT-GTTCTAGCTGGTGCTATCACTATACTTTTAACAGATC-GAAACTTAAATACATCTTTCTTTGACCCAGCCGGAGG-GGGAGATCCAATCCTTTACCAACATTTATTT

To submit an ASV table for upload, users must complete a small number of steps. First, the project container which will later hold the uploaded table must be selected. This is necessary for later access management to the uploaded data. Then, after selecting the desired table file, its layout will be automatically analysed and those columns that will be read into the database will be highlighted. In case this automatic highlighting fails, the user will be prompted to manually select the columns for each data type. If the table in question is a new version of a table that already exists in the ASV Registry, this table can be selected as the parent table. Finally, a set of sequencing primers can be added or chosen from a list of previously used primers. Primer declaration is recommended, but not obligatory. This allows for the upload of datasets for which the used primers are not known or should not be disclosed. Once uploaded, the table receives a unique persistent identifier, in the form of a simple integer, by which it will be findable at any later point. Each sequence within the table will also be saved to the database as part of an ASV set. An ASV set consists of the sequence and the sequencing primer pair selected during the table upload and is assigned a unique identifier – called ASV Set ID – in the format

GBOL_ASV_ID_X

where X corresponds to a continuously incremented integer (Table 2, yellow column). The combination of sequence data and corresponding sequencing primers allows for more precise data points while limiting redundancy. If there have been no primers declared for a given sequence, this will still result in the assignment of a distinct ASV Set ID. As a result, this set will behave identically to any set for which primers were declared.

It is worth noting that the assignment of ASV Set ID only takes the exact sequence into consideration. This means that two sequences will be assigned the

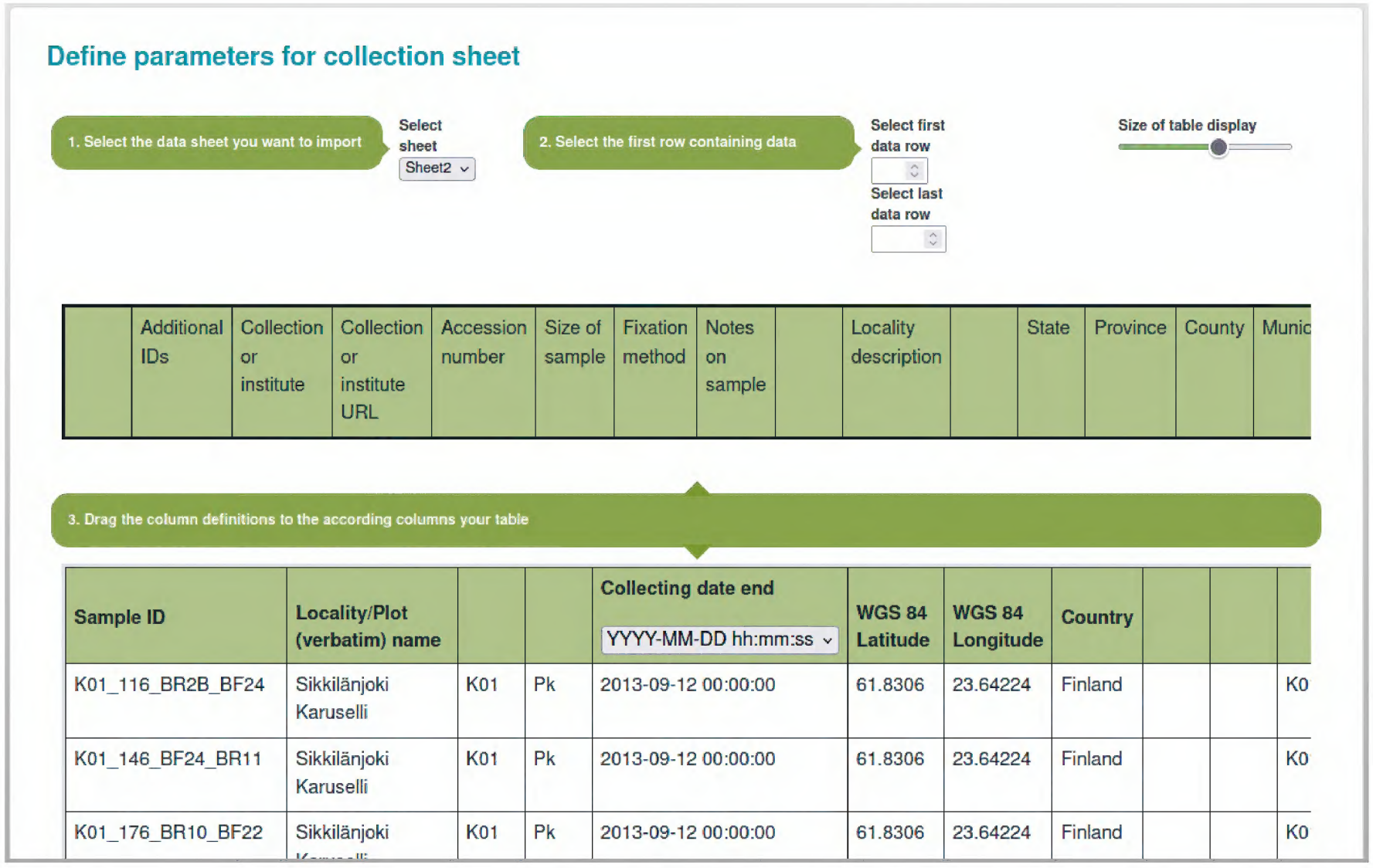

Figure 2. Collection table upload. Users can and should upload collection metadata for all uploaded ASV tables. The upload of these collection metadata does not require a dedicated format. Rather, users must assign an attribute to each column in the metadata table per drag-and-drop. This compromise allows for the easiest handling of this diverse data by both the user and the ASV Registry.

same identifier only if they match each other exactly and the selected primers are identical or no primers were selected. Consequently, a length discrepancy between two otherwise identical sequences will result in the registration of two separate ASV sequences in the ASV Registry. The registration of sequence subsets, based on different sequence lengths, is thus not possible. We realise that this may limit the comprehensiveness or meaningfulness of data in some cases; however, this approach allows sequences and corresponding annotations to be compared with a large degree of independence from the method or pipeline used to generate the respective sequences or datasets.

The ASV Registry also provides a versioning utility. When uploading an ASV table, users may choose to upload that table as a new version of a previously uploaded table. If this is the case, the lineage is noted within the database and allows users to create a series of tables. Such a series facilitates keeping track of changes made to the original data, as may be useful when testing different workflows or pipelines. Each version can be independently annotated, potentially highlighting differences between such datasets in terms of the discovered species.

Despite our registration approach and the versioning utility, we must still acknowledge that the comparability of datasets that have not been produced using the same pipeline is subject to various potential sources of bias and underlies the user's scrutiny. During or before table publication, users may add links to raw read data on INSDC (International Nucleotide Sequence Database Collaboration). For this, the URI and accession number of each dataset of interest must be provided. We encourage users to make use of this to enhance the reproducibility of their datasets.

Table 2. Assignment of persistent identifiers during upload. Each individual ASV uploaded as part of an ASV table receives a unique, persistent identifier called ASV Set ID (yellow; see Table 1 for the other colour codes). Notably, the unique combination of sequence and primers selected during table upload constitutes a new ASV set. This approach to ID assignment greatly limits redundancy within the database and helps to make recurring amplicons evident.

asv_id	ASV_Set_ID	PCRPCR1C	PCRPCR1a	PCRPCR1b	PCRPCR2C	PCRPCR2a	PCRPCR2b	Seq
uniq152440	GBOL_ASV_ID_97	1	3	1	0	1	0	TCTATCTTCATACTTATTTCATTCTTCTC-CATCAATTGATATTGCAATTTTTTCAATA-CATATTACAGGAATTTCCTCTGTTATTG-GTTCTTTAAATTTTATTGTAACTAT-TATATTTATAAAAAATTTTTCCTTAAAT-TATGATCAAATCAATTTATTTTCTT-GATCTGTATGTATTACAGTAATTC-TACTAATTTTATCTTTACCAGTTTTAG-CAGGAGCAGTAACTATACTTCTATTT-GATCGAAATTTTAATACATCCTTTTTT-GATCCTATAGGAGGAGGTGATCCAAT-TCTTTATCAACATTTATTC
uniq152443	GBOL_ASV_ID_41385	1	0	3	0	1	1	TTTATCATCTAGAATTGCTCATG-GAGGAGCTTCCGTAGATTTAGC-TATTTTCTCTCTTCACCTAGCAG-GAATTTCTTCAATTCTTGGGGCAG-TAAATTTTATTACAACTGTAATTA-ATATACGTTCTACAGGAATTACCTTT-GATCGAATACCTTTATTTGTAT-GATCTGTAGTTATTACAGCTATTTTAT-TACTTCTTTCTTTACCCGTATTAGCAG-GAGCAATTACTATATTATTAACAGATC-GAAATCTAAATACCTCATTCTTTGAC-CCAGCAGGAGGAGGAGACCCTATCT-TATATCAACATTTATTC
uniq152454	GBOL_ASV_ID_41390	3	1	0	0	2	0	ACTTTCATCTAATATTGCCCATG-GAGGTAGTTCTGTAGATCTTGC-TATTTTTTCTCTTCATTTAGCTG-GAATTTCTTCTATTCTAGGAGCTAT-TAATTTTATTACTACCATTATTA-ATATACGATTAAATAGTTTATCTTTT-GATCAAATACCTTTATTTATCTGAGCT-GTTGGAATTACAGCATTTTTATTATTAT-TATCTTTACCTGTTTTAGCAGGAGCTAT-TACTATATTATTAACTGATCGTAATTTA-AATACATCATTTTTTGACCCTGCAG-GAGGGGGAGACCCTATTTTATATCAA-CACCTATTT
uniq152461	GBOL_ASV_ID_41391	2	0	3	0	0	1	ACTATCAAATAATATATCCCATGCAG-GAGCATCAGTAGATTTAACAATTTTTTC-CCTTCACTTAGCCGGAATTTCATC-TATTCTAGGAGCTATTAATTTCATTTC-TACAGTAATTAATATACGAACTA-CAGGAATAAAATTTGAACGAATAC-CCTTATTTATTTGAGCTGTTACTTTA-ACAGCAATTCTATTATTATTATCAT-TACCAGTTTTAGCTGGAGCGATTA-CAATATTATTAACTGATCGAAACTTA-AATACTTCATTTTTTGATCCTTCAG-GTGGAGGAGACCCTATTTTATATCAA-CATTTATTT
uniq152464	GBOL_ASV_ID_41392	1	3	2	0	0	0	ACTTGCAGGAGCTATTGCTCATG-GTGGAGGATCAGTTGATCTAGC-CATTTTTTCATTACATCTAGCCGGT-GTCTCATCAATTTTAGGAGCAGTA-AACTTCATTACAACAGCAATCAACAT-GCGGTCGGAAAGAATAACACTGGAT-CAAACACCACTATTTGTTTGATCAG-TAGCAATCACAGCACTTCTATTAC-TACTATCATTACCAGTACTAGCAG-GAGCAATTACTATATTATTAACTGATC-GAAATTTAAATACATCATTCTTCGAC-CCAGCTGGAGGTGGAGATCCAAT-TCTCTATCAACATTTATTT

asv_id	ASV_Set_ID	PCRPCR1C	PCRPCR1a	PCRPCR1b	PCRPCR2C	PCRPCR2a	PCRPCR2b	Seq
uniq152471	GBOL_ASV_ID_41388	1	1	3	1	0	0	ATTGTCGAACAATATATTTCACAGAGGTGCCTCAGTAGACATAGCCATCTTTTCACTACACTTAGCAGGTGTATCATCTATCTTAGGAGCCATTAACTTTATTTCCACAATCATGAACATACGACCCACAGGTATAACACCTGAACAAATTCCCCTATTCGTATGATCTGTAGGAATCACTGCTCTATTATTATTATTATCACTCCCAGTTTTAGCCGGAGCAATCACTATATTATTAACAGACCGAAACTTCAACACATCCTTCTTTGACCCTACGGGGGGAGGAGATCCTATTCTATACCAACACTTATTT
uniq152480	GBOL_ASV_ID_41389	0	3	1	0	1	1	ACTATCTGCTAATATTGCTCATAGAGGAGCATCAGTAGATATAGCAATTTTCTCACTACATTTAGCAGGAGTATCATCAATTCTAGGAGCAGTAAACTTCATTTCAACTATTATTAATATACGGCCCGTAGGTATATCACCTGAACGAATTCCGCTATTTGTATGATCAGTAGGAATTACTGCATTACTATTATTATTATCACTACCTGTATTAGCAGGGCTATTACAATACTATTAACAGATCGTAATTTTAATACATCATTTTTTGACCCAGCCGGAGGTGGGGATCCTATTTTATATCAACATTTATTC
uniq152481	GBOL_ASV_ID_41384	0	3	1	1	1	0	TCTATCCTCAAATATTGCACATAACGGTTCATCTGTAGACCTGGCAATTTTCAGATTACATTTAGCGGGGATCTCCTCAATTCTAGGAGCAGTTAATTTTATTTCTTCTATAATTAATATACGACCTTTAAAAATAAGATTTGATCAAATACCTCTTTTTACTTGAGCTGTAGCTATTACAGCTCTCTTACTCCTCCTCTCTCTGCCCGTACTAGCTGGAGCGATTACTATACTTCTCACAGATCGAAACATTAATACATCATTTTTTGATCCCGCAGGAGGTGGAGATCCTATTCTATACCAACACTTATTT
uniq152482	GBOL_ASV_ID_41387	0	5	0	0	1	0	ACTTGCAGGAGCTATTGCTCATGGTGGGTGATCAGTAGATCTAGCCATTTTTACATTACATCTAGCCGGTGTTTCATCAATTTTAGGAGCAGTAAACTTTATTAAAACAGCAATTAATATGCGGTCGGGGAGAATAACACTGGATCAAACCACTATTTGTCTGATCAGTAGCAATTACAGCACTTTTATTACTATTACCATTACCAGTTTTAGCAGGAGCAATTACTATATTATTAACTGATCGAAACTTAAATACATCATTCTTTGATCCAGCTGGAGGTGGAGATCCAATTCTTTATCAACATTTATTC
uniq152483	GBOL_ASV_ID_41386	0	2	2	1	1	0	CCTATCTTCAGGAATTGCACATGGAGGTGCTTCTGTAGACTTAGCTATTTTTTCTCTACATTTAGCTGGTGTCTCATCAATTTTAGGAGCAGTAAATTTTATTACTACAGTTATTAATATACGATCTACTGGAATTACTTTCGATCGAATGCCATTGTTTGTATGAGCAGTGGTAATTACTGCTTTATTACTTCTCCTTTCTTTACCTGTTCTAGCTGGTGCTATCACTATACTTTTAACAGATCGAAACTTAAATACATCTTTCTTTGACCCAGCCGGAGGGGGAGATCCAATCCTTTACCAACATTTATTT

After the initial upload, the table enters the dedicated ASV Table life cycle and becomes enrichable in various ways: firstly, users can add collection metadata – referred to as Collection tables – to describe the sampling process and locations. Incorporation of these metadata greatly enhances the reusability, especially in the context of environmental monitoring. Hence, we encourage all users

to upload such metadata when uploading ASV tables. The collection metadata must be in XLSX format. It should be organised in a basic tabular fashion (see Fig. 2). We have deliberately chosen not to impose strict format requirements for these data due to the large number of possible metadata attributes that may be included in these Collection tables. Users must assign the corresponding attribute to each column in the table from a list of 26 available metadata attributes, like habitat description, coordinates, fixation methods, collection date and size of sample. Once uploaded, the location data from the collection table will be linked to each ASV sequence within the database. This allows the sequences to be displayed and searched geographically on the map of the dedicated search pages.

Further, the sequences in the table can receive taxonomic assignments. The registry provides thirteen reference databases for this purpose, including multiple SILVA (Quast et al. 2013), NCBI (https://www.ncbi.nlm.nih.gov/) and PLANiTS (Banchi et al. 2020) datasets, respectively. Users can set various parameters for this process and also select whether the chosen databases should be queried in parallel or sequentially. When annotating a single ASV sequence, all hits returned by the respective algorithm are stored in the database together with the taxon found, the associated taxonomy, E-value and percent identity. With the default settings, VSEARCH only returns the best hit, while the BLAST and BOLD algorithms return lists of hits. Regardless of the number of hits returned, for a given ASV sequence and reference database, only the taxon identified by the last available hit is displayed in the ASV registry.

The specimen in the respective reference database, identified for a given sequence, is also saved by its ID and accessible with an enclosed link. This way, further information or metadata held by the reference databases are easily accessible and may supplement the information provided by the taxonomic annotation. All of these data are backtracked to the individual ASV sequences within the database (see Appendix 1: Fig. A1). As the assignments are subject to change over time, repeated taxonomic annotation is supported by the ASV Registry, thereby allowing users to keep their assignments up to date. For this reason, only the most recent taxon annotation from a specific reference database is displayed for an ASV sequence. Even if all hits from all annotation runs are stored in the database, we assume that the most recent annotation is the most accurate, as the quality of the reference database increases over time (Weigand et al. 2019). Any assignments found will be attached to both tables and individual sequences within the database (see Appendix 1: Fig. A1). Methods and parameters of the respective assignment runs are also tracked.

ASV tables are available for download immediately after upload. If no taxonomic assignments have been performed on a table, the download will return the original file, only annotated with the unique identifier assigned to each ASV sequence. If a table were taxonomically annotated, users can tailor which annotation data to include and where and exclude data of no interest (Fig. 3). This way, taxonomy data from several runs and over time can be readily compared in the downloaded file (Table 3). Users can also choose in which format to download the table to accommodate potential downstream applications; CSV, TSV and XLSX formats are available. Due to the potential size of the uploaded and analysed tables, download preparation is managed by a queue to allow users to continue using the ASV Registry during processing. Users are notified by email once processing is complete and are provided with a link to access the requested file.

Table 3. Taxonomic annotation. During the taxonomic annotation process, various assignment data are recorded and saved. Users can then decide which of these data to include, in which order to arrange data from individual annotation processes and where within the final table to include these data. This allows the final ASV table to be maximally informative without becoming chaotic and hard to read. In this case, two separate databases have been queried (pink and purple) and, in each case, the found taxa and percentages of identity have been added to the downloaded table file. For the other colour codings, see Table 1.

		Blast run 69 from 20240124-12 h15 m		Blast run 69 from 20240124-12 h15 m								
		GBOL		BOLD								
asv_id	ASV_Set_ID	Taxon Name	Percent Identity	Taxon Name	Percent Identity	PCRPCR1C	PCRPCR1a	PCRPCR1b	PCRPCR2C	PCRPCR2a	PCRPCR2b	Seq
uniq152440	GBOL_ASV_ID_97	Bombus lapidarius	100	Bombus lapidarius atlanticus	100	1	3	1	0	1	0	TCTATCTTCATACTTATTTCATTCTTCTCCATCAATTGATATTGCAATTTTTTCAATACATATTACAGGAATTTCCTCTGTTATTGGTTCTTTAAATTTTATTGTAACTATTATATTTATAAAAAATTTTTCCTTAAATTATGATCAAATCAATTTATTTTCTTGATCTGTATGTATTACAGTAATTCTACTAATTTTATCTTTACCAGTTTTAGCAGGAGCAGTAACTATACTTCTATTTGATCGAAATTTTAATACATCCTTTTTTGATCCTATAGGAGGAGGTGATCCCAATTCTTTATCAACATTTATTC
uniq152443	GBOL_ASV_ID_41385	Empis tessellata	100	Empis tessellata	100	1	0	3	0	1	1	TTTATCATCTAGAATTGCTCATGGAGGAGCTTCCGTAGATTTAGCTATTTTCTCTCTTCACCTAGCAGGAATTTCTTCAATTCTTGGGGCAGTAAATTTTATTACAACTGTAATTAATATACGTTCTACAGGAATTACCTTTGATCGAATACCTTTATTTGTATGATCTGTAGTTATTACAGCTATTTTATTACTTCTTTCTTTACCCGTATTAGCAGGAGCAATTACTATATTATTAACAGATCGAAATCTAAATACCTCATTCTTTGACCCAGCAGGAGGAGGAGACCCTATCTTATATCAACATTTATTC
uniq152454	GBOL_ASV_ID_41390	Axylia putris	100	Axylia putris	100	3	1	0	0	2	0	ACTTTCATCTAATATTGCCCATGGAGGTAGTTCTGTAGATCTTGCTATTTTTTCTCTTCATTTAGCTGGAATTTCTTCTATTCTAGGAGCTATTAATTTTATTACTACCATTATTAATATACGATTAAATAGTTTATCTTTTGATCAAATACCTTTATTTATCTGAGCTGTTGGAATTACAGCATTTTTATTATTATTATCTTTACCTGTTTTAGCAGGAGCTATTACTATATTATTAACTGATCGTAATTTAAATACATCATTTTTTGACCCTGCAGGAGGGGGAGACCCTATTTTATATCAACACCTATTT
uniq152461	GBOL_ASV_ID_41391	Macrophya diversipes	100	Macrophya diversipes	100	2	0	3	0	0	1	ACTATCAAATAATATATCCCATGCAGGAGCATCAGTAGATTTAACAATTTTTTCCCTTCACTTAGCCGGAATTTCATCTATTCTAGGAGCTATTAATTTCATTTCTACAGTAATTAATATACGAACTACAGGAATAAAATTTGAACGAATACCCTTATTTATTTGAGCTGTTACTTTAACAGCAATTCTATTATTATTATCATTACCAGTTTTAGCTGGAGCGATTACAATATTATTAACTGATCGAAACTTAAATACTTCATTTTTTGATCCTTCAGGTGGAGGAGACCCTATTTTATATCAACATTTATTT

		Blast run 69 from 20240124-12 h15 m		Blast run 69 from 20240124-12 h15 m							
		GBOL		BOLD							
asv_id	ASV_Set_ID	Taxon Name	Percent Identity	Taxon Name	Percent Identity	PCRPCR1C	PCRPCR1a	PCRPCR1b	PCRPCR2C	PCRPCR2a	PCRPCR2b	Seq
uniq152464	GBOL_ASV_ID_41392			Omocestus petraeus	99	1	3	2	0	0	0	ACTTGCAGGAGCTATTGCTCATGGTGGAGGATCAGTTGATCTAG-CCATTTTTTCATTACATCTAGCCGGTGTCTCATCAATTTTAG-GAGCAGTAAACTTCATTACAACAGCAATCAACATGCGGTCG-GAAAGAATAACACTGGATCAAACACCACTATTTGTTTGATCAG-TAGCAATCACAGCACTTCTATTACTACTATCATTACCAGTACTAG-CAGGAGCAATTACTATATTATTAACTGATCGAAATTTAAATACAT-CATTCTTCGACCCAGCTGGAGGTGGAGATCCAATTCTCTATCAA-CATTTATTT
uniq152471	GBOL_ASV_ID_41388	Gastrodes abietum	100			1	1	3	1	0	0	ATTGTCGAACAATATATTTCACAGAGGTGCCTCAGTAGACATAGC-CATCTTTTCACTACACTTAGCAGGTGTATCATCTATCTTAGGAGC-CATTAACTTTATTTCCACAATCATGAACATACGACCCACAGGTATA-ACACCTGAACAAATTCCCCTATTCGTATGATCTGTAGGAATCACT-GCTCTATTATTATTATTATCACTCCCAGTTTTAGCCGGAGCAAT-CACTATATTATTAACAGACCGAAACTTCAACACATCCTTCTTTGAC-CCTACGGGGGGAGGAGATCCTATTCTATACCAACACTTATTT
uniq152480	GBOL_ASV_ID_41389	Himacerus apterus	100	Himacerus apterus	99	0	3	1	0	1	1	ACTATCTGCTAATATTGCTCATAGAGGAGCATCAG-TAGATATAGCAATTTTCTCACTACATTTAGCAGGAGTATCATCAAT-TCTAGGAGCAGTAAACTTCATTTCAACTATTATTAATATACGGC-CCGTAGGTATATCACCTGAACGAATTCCGCTATTTGTATGATCAG-TAGGAATTACTGCATTACTATTATTATTATCACTACCTGTATTAG-CAGGGGCTATTACAATACTATTAACAGATCGTAATTTTAATACAT-CATTTTTTGACCCAGCCGGAGGTGGGGATCCTATTTTATATCAA-CATTTATTC
uniq152481	GBOL_ASV_ID_41384	Attelabus nitens	100			0	3	1	1	1	0	TCTATCCTCAAATATTGCACATAACGGTTCATCTGTAGACCTGG-CAATTTTCAGATTACATTTAGCGGGGATCTCCTCAATTCTAGGAG-CAGTTAATTTTATTTCTTCTATAATTAATATACGACCTTTAAAAATA-AGATTTGATCAAATACCTCTTTTTACTTGAGCTGTAGCTATTA-CAGCTCTCTTACTCCTCCTCTCTCTGCCCGTACTAGCTGGAGC-GATTACTATACTTCTCACAGATCGAAACATTAATACATCATTTTTT-GATCCCGCAGGAGGTGGAGATCCTATTCTATACCAACACTTATTT
uniq152482	GBOL_ASV_ID_41387					0	5	0	0	1	0	ACTTGCAGGAGCTATTGCTCATGGTGGGTGATCAGTAGATCTAG-CCATTTTTACATTACATCTAGCCGGTGTTTCATCAATTTTAG-GAGCAGTAAACTTTATTAAAACAGCAATTAATATGCGGTCGG-GGAGAATAACACTGGATCAAACCACTATTTGTCTGATCAGTAG-CAATTACAGCACTTTTATTACTATTACCATTACCAGTTTTAG-CAGGAGCAATTACTATATTATTAACTGATCGAAACTTAAATACAT-CATTCTTTGATCCAGCTGGAGGTGGAGATCCAATTCTTTATCAA-CATTTATTC
uniq152483	GBOL_ASV_ID_41386	Nemopoda nitidula	100	Nemopoda nitidula	100	0	2	2	1	1	0	CCTATCTTCAGGAATTGCACATGGAGGTGCTTCTGTAGACTTAGC-TATTTTTTCTCTACATTTAGCTGGTGTCTCATCAATTTTAGGAG-CAGTAAATTTTATTACTACAGTTATTAATATACGATCTACTGGAAT-TACTTTCGATCGAATGCCATTGTTTGTATGAGCAGTGGTAATTACT-GCTTTATTACTTCTCCTTTCTTTACCTGTTCTAGCTGGTGCTAT-CA CTATACTTTTAACAGATCGAAACTTAAATACATCTTTCTTTGAC-CCAGCCGGAGGGGGAGATCCAATCCTTTACCAACATTTATTT

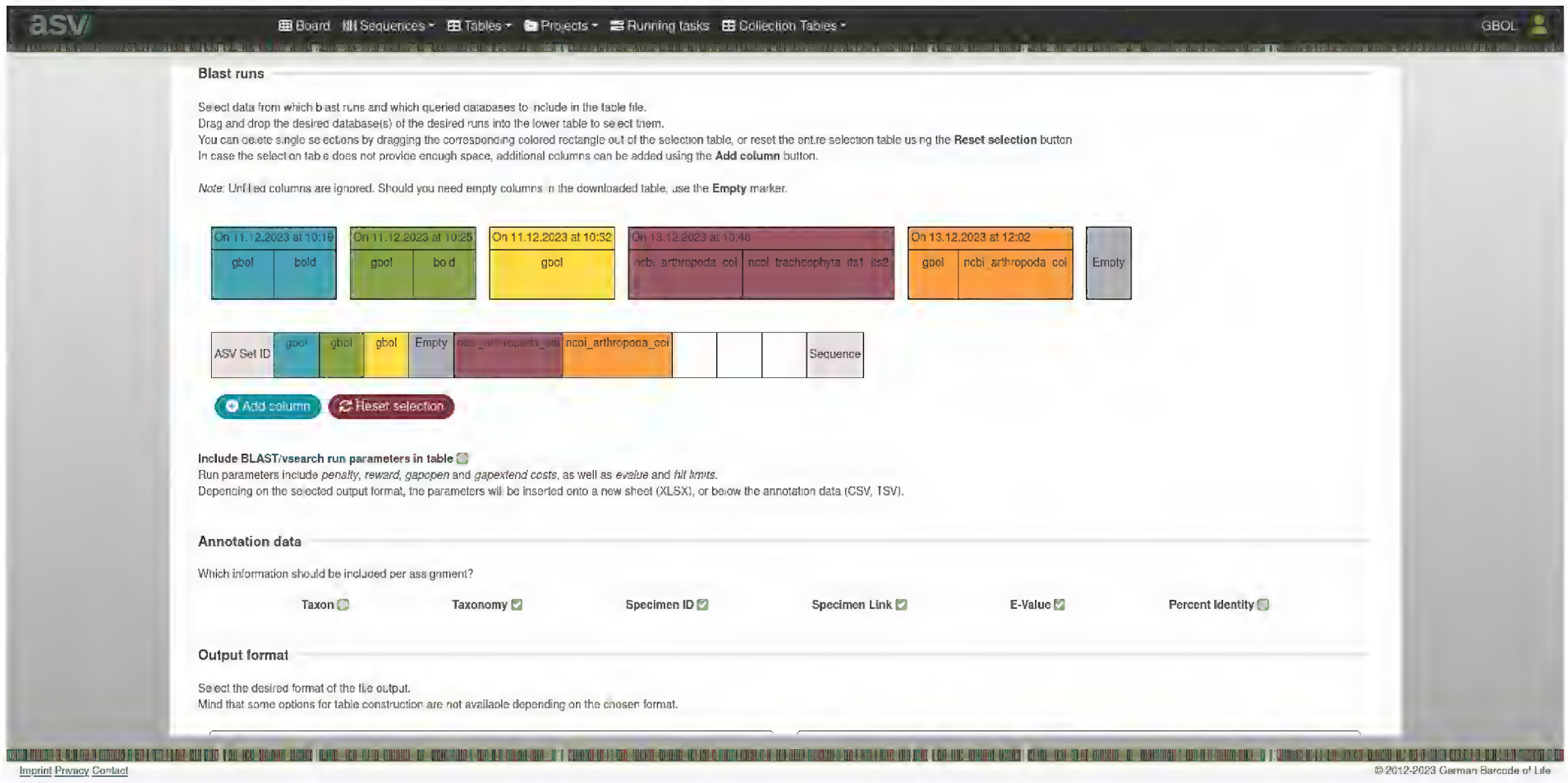
asv
Board
Sequences
Tables
Projects
Running tasks
Collection Tables
GBOL
Blast runs
Select data from which blast runs and which queried databases to include in the table file.
Drag and drop the desired database(s) of the desired runs into the lower table to select them.
You can delete single selections by dragging the corresponding colored rectangle out of the selection table, or reset the entire selection table using the Reset selection button
In case the selection table does not provide enough space, additional columns can be added using the Add column button.
Note: Unfilled columns are ignored. Should you need empty columns in the downloaded table, use the Empty marker.
On 11.12.2023 at 10:19
gbol
bold
On 11.12.2023 at 10:25
gbol
bold
On 11.12.2023 at 10:32
gbol
On 13.12.2023 at 10:48
ncbi_arthropoda_coi
ncbi_tracheophyta_its1_its2
On 13.12.2023 at 12:02
gbol
ncbi_arthropoda_coi
Empty
ASV Set ID
gbol
gbol
gbol
Empty
ncbi_arthropoda_coi
Sequence
Add column
Reset selection
Include BLAST/vsearch run parameters in table
Run parameters include penalty, reward, gapopen and gapextend costs, as well as evalue and hit limits.
Depending on the selected output format, the parameters will be inserted onto a new sheet (XLSX), or below the annotation data (CSV, TSV).
Annotation data
Which information should be included per assignment?
Taxon
Taxonomy
Specimen ID
Specimen Link
E-Value
Percent Identity
Output format
Select the desired format of the file output.
Mind that some options for table construction are not available depending on the chosen format.
Imprint Privacy Contact
© 2012-2023 German Barcode of Life

Figure 3. Table download options. Users have numerous options for tailoring downloaded tables to their needs. It is possible to choose queried databases from selected annotation runs, respectively. The order in which annotation runs appear and the data points to be shown for each annotation are also selectable. The file format for the output file can also be chosen.

ASV sequences can also be uploaded separately, although the upload and management of ASV tables is the core of the ASV Registry. Sequences must be provided in FASTA format, either in the form of a file or by copy-and-pasting into the dedicated ASV Sequence Upload page of the ASV Registry. A project within which the sequences will be uploaded must be chosen and a sequencing primer pair may be selected. Just as ASV sequences that are uploaded as part of an ASV table, these sequences will be assigned an ASV Set ID and saved in the database; however, taxonomic annotation and linkage of collection metadata are only possible for ASV tables. Hence, if a sequence – or rather an ASV set – does not also appear in an ASV table, it cannot be further enriched. Individual sequences or datasets consisting exclusively of sequences cannot be published in the ASV Registry. We therefore encourage all users to upload their data as ASV tables if possible to benefit from all functions of the ASV Registry.

Once satisfied with the state of a table, users can then choose to publish their table. First, a publication draft is created for the table and a preliminary DOI is assigned that can already be used to point to the table within the ASV Registry. Users must provide a number of metadata for the ASV table at hand before it can be submitted for publication. A title and short abstract for the table must be given, as well as full name and affiliation of one or more authors. Further, the declaration of a licence is obligatory. We recommend the use of a CC BY licence in order to foster unrestricted use of the published datasets. The very nature of the ASV Registry requires the uploaded datasets to allow derivation and dissemination. In the case that a table, that is already part of an existing publication, is to be published via the ASV Registry, the licence granted by the original publisher should be declared here. Apart from this obligatory information, a number of optional metadata may be provided to enrich the

published table or improve findability. Methods like collection and processing parameters can be noted, keywords and links to INSDC datasets may be added, as well as citations for related publications. The sum of these metadata is submitted to DataCite (https://datacite.org/) and the preliminary DOI is finalised. This moves the table out of its original project container and access is, thus, no longer limited to project members (Fig. 1). Instead, it becomes available to all users – logged in or not – to view, download and use.

The ASV Registry provides several methods to search through and retrieve project-internal and published data. These include, but are not limited to, filtering sequences by taxon assignment, tables and sequences by project and sequences by location in case collection tables were uploaded (Fig. 4).

In the future, we plan to position the ASV Registry as a contributor to the global metabarcoding data infrastructure. The aim is to connect the uploaded occurrence data via the GBIF (Global Biodiversity Information Facility) provider LIB (Leibniz Institute for the Analysis of Biodiversity Change) and to implement the GBIF metabarcoding pipeline for automatic provision of uploaded occurrence data to GBIF.

Figure 4. Table search page. Published tables and not yet published tables from the signed-in user's projects can be easily searched and retrieved on the dedicated web page. Aside from filters and a map, an option for keyword search featuring different modes and input suggestions also exists.

Conclusions

The ASV Registry, the web application we present here, aims to pave the way for efficient and effective use of ASV data, which is rarely possible at the moment. By providing a platform for researchers to archive, analyse and ultimately publish their ASV data, we hope to foster comparative and collaborative DNA metabarcoding research and biodiversity monitoring in the future. Easier access to published data and structured publication of new data will certainly boost the collective usefulness of DNA metabarcoding.

Project description

The presented software has been developed within a dedicated work package of the GBOL III: Dark Taxa project. The registry builds upon and makes use of the work done in the previous two funding phases of the German Barcode of Life (GBOL) project to establish a barcode library for the flora, fauna and fungi of Germany while supporting the current research to investigate so-called dark taxa – species-rich, but so far understudied groups of insects (Hausmann et al. 2020; Rduch and Peters 2020).
Study area description: Metabarcoding, metagenomics and bioinformatics
For further information, see https://gbol.bolgermany.de/gbol3/de/gbol-dark-taxa/.
Web location (URIs): The ASV data portal is available at https://asv.bolgermany.de/metabarcoding.
The source code is available at: https://gitlab.leibniz-lib.de/GBOL/asv-registry.
Technical specification: Platform: web application
Programming language: Python, Javascript, MySQL
Operational system: various
Usage rights: The source code is subject to an MIT Licence. Usage for private or scientific purposes must comply with the terms of this licence.

Usage of data published in the ASV data portal must comply with the terms of the licence chosen by the user that holds ownership of the respective data.

Acknowledgements

Further, we are grateful to Dr. Ralph Peters and all other members of the GBOL III: Dark Taxa project at the Leibniz-Institute for the Analysis of Biodiversity Change for their ongoing support during code development and testing.

We would also like to thank Dr. Tamara Hartke, Dr. Vera Zizka, Vera Prenzel and Hadeel Ragab for kindly allowing us to use data from their projects for illustrative purposes.

Additional information

Conflict of interest

The authors have declared that no competing interests exist.

Ethical statement

No ethical statement was reported.

Funding

This research was funded by the Federal Ministry of Education and Research of Germany (Bundesministerium für Bildung und Forschung, BMBF) as part of the project “GBOL III: Dark Taxa” as Research for Sustainable Development (Forschung für Nachhaltige Entwicklung, FONA; www.fona.de) under the funding reference 16LI1901A.

Author contributions

Conceptualization: BQ, PG. Software: SM, CB, BQ. Supervision: PG, BQ. Writing - original draft: SM, CB. Writing - review and editing: BQ, CB, VR, PG, SM.

Author ORCIDs

Sandra Meid https://orcid.org/0000-0003-4627-1853
Björn Quast https://orcid.org/0000-0002-3760-5834
Vera Rduch https://orcid.org/0000-0002-6499-2876
Peter Grobe https://orcid.org/0000-0003-4991-5781

Data availability

All of the data that support the findings of this study are available in the main text.

References

Antich A, Palacin C, Wangensteen OS, Turon X (2021) To denoise or to cluster, that is not the question: Optimizing pipelines for COI metabarcoding and metaphylogeography. BMC Bioinformatics 22(177): 177. https://doi.org/10.1186/s12859-021-04115-6

Banchi E, Ametrano CG, Greco S, Stanković D, Muggia L, Pallavicini A (2020) PLANiTS: A curated sequence reference dataset for plant ITS DNA metabarcoding. Database (Oxford) 2020: baz155. https://doi.org/10.1093/database/baz155

Blaxter M, Mann J, Chapman T, Thomas F, Whitton C, Floyd R, Abebe E (2005) Defining operational taxonomic units using DNA barcode data. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 360(1462): 1935–1943. https://doi.org/10.1098/rstb.2005.1725

Callahan BJ, McMurdie PJ, Holmes SP (2017) Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME Journal 11(12): 2639–2643. https://doi.org/10.1038/ismej.2017.119

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: Architecture and applications. BMC Bioinformatics 10(1): 421. https://doi.org/10.1186/1471-2105-10-421

Elbrecht V, Vamos EE, Meissner K, Aroviita J, Leese F (2017) Assessing strengths and weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream monitoring. Methods in Ecology and Evolution 8(10): 1265–1275. https://doi.org/10.1111/2041-210X.12789

German Barcode of Life Consortium [Wägele W, Haszprunar G, Eder J, Xylander W, Borsch T, Quandt D, Grobe P, Pietsch S, Geiger M, Astrin J, Rulik B, Hausmann A, Moriniere J, Holstein J, Krogmann L, Monje C, Traunspurger W, Hohberg K, Lehmitz R, Müller K, Nebel M, Hand R] (2011) GBOL Webportal. [Dataset]. Version: 20170316. Data Publisher: Zoological Research Museum Koenig - Leibniz Institute for Animal Biodiversity. https://www.bolgermany.de

Griffin PC, Khadake J, LeMay KS, Lewis SE, Orchard S, Pask A, Pope B, Roessner U, Russell K, Seemann T, Treloar A, Tyagi S, Christiansen JH, Dayalan S, Gladman S, Hangartner SB, Hayden HL, Ho WWH, Keeble-Gagnère G, Korhonen PK, Neish P, Prestes PR, Richardson MF, Watson-Haigh NS, Wyres KL, Young ND, Schneider MV (2017) Best practice data life cycle approaches for the life sciences. F1000 Research 6: 1618. https://doi.org/10.12688/f1000research.12344.1

Hausmann A, Krogmann L, Peters RS, Rduch V, Schmidt S (2020) GBOL III: Dark Taxa. iBOL Barcode Bulletin 10(1): 1–4. https://doi.org/10.21083/ibol.v10i1.6242

Leese F, Woppowa L, Bálint M, Höss S, Krehenwinkel H, Lötters S, Meissner K, Nowak C, Rausch P, Rduch V, Rulik B, Weigand AM, Zimmermann J, Koschorrek J, Zügh art W (2023) DNA-basierte Biodiversitätsanalysen im Natur- und Umweltschutz: Welche

Optionen haben wir für eine Standardisierung? Eine Handlungsempfehlung aus Forschung & Praxis. BfN Schriften 666, 53 pp. https://doi.org/10.19217/skr666

Macher TH, Beermann AJ, Leese F (2021) TaxonTableTools: A comprehensive, platform-independent graphical user interface software to explore and visualise DNA metabarcoding data. Molecular Ecology Resources 21(5): 1705–1714. https://doi.org/10.1111/1755-0998.13358

Mitchell AL, Scheremetjew M, Denise H, Potter S, Tarkowska A, Qureshi M, Salazar GA, Pesseat S, Boland MA, Hunter F, ten Hoopen P, Alako B, Amid C, Wilkinson DJ, Curtis TP, Cochrane G, Finn RD (2017) EBI Metagenomics in 2017: Enriching the analysis of microbial communities, from sequence reads to assemblies. Nucleic Acids Research 46(D1): D726–D735. https://doi.org/10.1093/nar/gkx967

Porter T, Hajibabaei M (2018) Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. Molecular Ecology 27(2): 313–338. https://doi.org/10.1111/mec.14478

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2013) The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Research 41(D1): D590–D596. https://doi.org/10.1093/nar/gks1219

Ratnasingham S, Hebert PDN (2007) BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). Molecular Ecology Notes 7(3): 355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

Rduch V, Peters RS (2020) GBOL III: Dark Taxa – die dritte Phase der German Barcode of Life Initiative hat begonnen. Koenigiana 14(2): 91–107.

Remmel N, Buchner D, Enss J, Hartung V, Leese F, Welti EAR, Sinclair JS, Haase P (2024) DNA metabarcoding and morphological identification reveal similar richness, taxonomic composition and body size patterns among flying insect communities. Insect Conservation and Diversity 17(3): 1–15. https://doi.org/10.1111/icad.12710

Richardson LJ, Allen B, Baldi G, Beracochea M, Bileschi M, Burdett T, Burgin J, Caballero-Pérez J, Cochrane G, Colwell L, Curtis T, Escobar-Zepeda A, Gurbich T, Kale V, Korobeynikov A, Raj S, Rogers AB, Sakharova E, Sanchez S, Wilkinson D, Finn RD (2023) MGnify: The microbiome sequence data analysis resource. Nucleic Acids Research 51(D1): D753–D759. https://doi.org/10.1093/nar/gkac1080

Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016) VSEARCH: a versatile open source tool for metagenomics. PeerJ 4: e2584. https://doi.org/10.7717/peerj.2584

Rüegg J, Gries C, Bond-Lamberty B, Bowen GJ, Felzer BS, McIntyre NE, Soranno PA, Vanderbilt KL, Weathers KC (2014), Completing the data life cycle: using information management in macrosystems ecology research. Frontiers in Ecology and the Environment 12: 24–30. https://doi.org/10.1890/120375

Weigand H, Beermann A, Čiampor F, Costa F, Csabai Z, Duarte S, Geiger M, Grabowski M, Rimet F, Rulik B, Strand M, Szucsich N, Weigand A, Willassen E, Wyler S, Bouchez A, Borja A, Čiamporová-Zaťovičová Z, Ferreira S, Dijkstra K, Eisendle U, Freyhof J, Gadawski P, Graf W, Haegerbaeumer A, van der Hoorn B, Japoshvili B, Keresztes L, Keskin E, Leese F, Macher J, Mamos T, Paz G, Pešić V, Pfannkuchen DM, Pfannkuchen MA, Price B, Rinkevich B, Teixeira ML, Várbíró G, Ekrem T (2019) DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. The Science of the Total Environment 678: 499–524. https://doi.org/10.1016/j.scitotenv.2019.04.247

Wilkinson M, Dumontier M, Aalbersberg I, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3(1): 160018. https://doi.org/10.1038/sdata.2016.18

Appendix 1. Database scheme

All ASV Registry data is stored in a relational database, the schema of which is presented and explained here (Fig. A1).

The database schema contains four higher-level blocks: the ASV table, the taxonomic assignment, the sequence registry and the project-specific metadata.

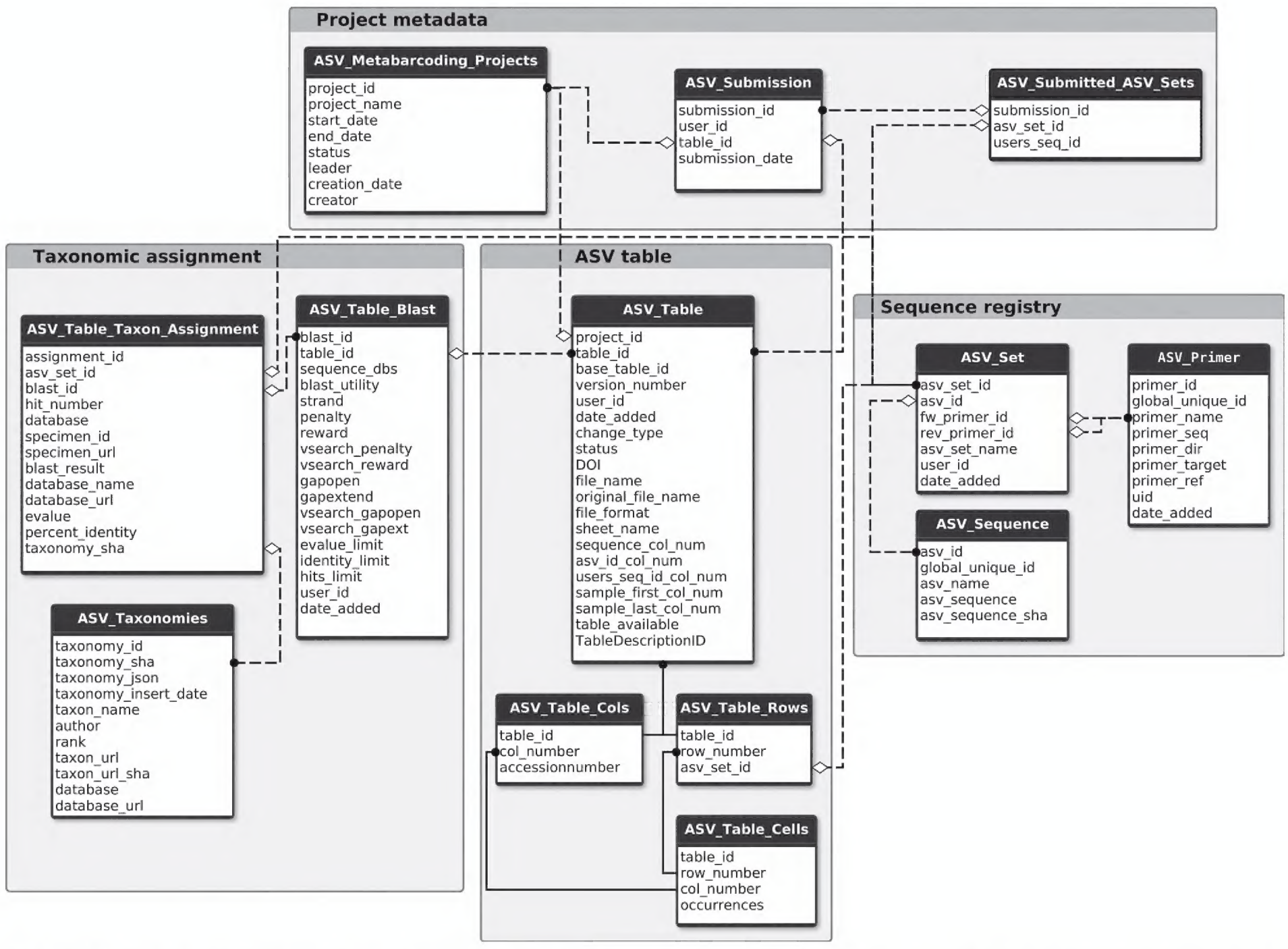

Figure A1. Relational database of the ASV Registry. During upload of an ASV table, the contained data are saved into a number of linked tables within the database. The ASV set – the combination of a single ASV sequence and the primer pair used for amplification – is of central importance and acts as a link between various database tables. Please note that this scheme is not exhaustive, but rather includes a limited number of tables for visualisation.

Project metadata: Uploaded ASV tables are assigned to exactly one project. Registered users have access to one or more projects (not shown in Fig. A1). The metadata for the uploads and their link to the sets of ASV tables are also managed in this block.

Sequence Registry: Unique ASV set IDs are generated via the combination of the primer set and the sequence information. Each ASV set is linked to a sequence and a taxonomic assignment. Unique ASV IDs are assigned during table upload when the data are saved to the database: here, the primary key asv_set_id is generated in the ASV_Set table and entered in the asv_set_name field with the prefix GBOL_ASV_ID_X (X stands for the integer from the primary key). ASV sequences can be uploaded by themselves as a fasta file without or as part of an ASV table. The latter is the preferred way. Each upload is tracked in the database table ASV_Submission with the date and time of the upload and the user ID.

ASV table: The uploaded ASV table, consisting of rows, columns and cells, is mapped in the relational database schema by the four tables with information about the entire table, its rows, columns and cells in the ASV table block, thus storing the original data loaded by the user. Each row of the loaded ASV table is uniquely assigned to an ASV set. This makes the block the heart of the registry for all operations on the loaded tables, for example, for the publication of tables or taxonomic assignment, as it is the connecting element between ASV sets, project and associated ASV sequences.

Subsequent changes of ASV tables are tracked in the database table ASV_History (not shown in Fig. A1). Users can either provide chains of subsequent versions or different versions to the same ASV table. In this way, working on different branches coming from the same origin is possible.

Taxonomic assignment: The assignment of taxon names to the sequences is saved in the taxonomic assignment block. All information on the search parameters used, the reference databases etc., as well as all information on the assigned taxon can be found here. The results of such an assignment are stored in a JSON structure within the database table ASV_Taxonomies, as the available reference databases use different ranks and hierarchies in their taxonomies. It is necessary to keep this information together to document the complete outcome of the assignment at a point in time and not to overwrite it with a generic taxonomy that might not fit the users' requirements.

ASV table Publication (not shown in Fig. A1): The database also stores metadata on the ASV tables such as authors, citations, previous or related publications and keywords. At the latest when a table becomes published via the ASV Registry, a title, an abstract and a licence for re-use must be added to the metadata. We force the use of a licence that allows derivations, as users must be able to make new taxon assignments in the registry. The metadata are also registered at datacite.org when the ASV table is published and a DOI is assigned to the publication.

Material and Methods: Protocols on DNA extraction, amplification and purification, as well as software pipelines are stored in the database table ASV_TableMethodsDescription (not shown in Fig. A1). Since the protocols and methods can vary over a wide range, we have decided to store this information in a single text field for each ASV table.